Are cognitive principles useful in data mining? A case study in unsupervised learning

Luis Talavera*

Abstract—**In the early days of machine learning, much work was motivated by concerns with human behavior. However, more recently, attention has shifted to the application of methods to real world problems as exemplified by new disciplines like data mining. In this paper we claim that cognitive based biases can still be useful for solving issues arising in real world problems. We present an example in unsupervised learning comparing the notion of selective attention in cognitive psychology with the problem of feature selection in machine learning and data mining. We show how a system that selects salient features during learning is able to fit psychological data and, at the same time, provide more efficient and comprehensible results with real world data. The conclusion is that the notion of bounded rationality or cognitive economy present in earlier research in machine learning, can actually be reinterpreted to help in solving data mining problems.**

I. Introduction

In the early days of machine learning, much work was motivated by concerns with human behavior. However, as the field has matured and showed itself capable of solving challenging real world problems, the focus on cognition has significantly decreased. Probably, there are two main trends in the nature of machine learning research that have contributed to this fact. First, the gap between machine learning and statistical research has narrowed, lessening the different biases that use to characterize both fields. Historically, statistics are more concerned with algorithmic and numerical methods, with machine learning rather focusing on symbolic problems and heuristic solutions. Nowadays, a great extent of machine learning research is turning to statistical concepts with a sound theoretical background. Secondly, the number of fielded applications of machine learning has increased rapidly in recent years. This shift towards an application-oriented view is particularly emphasized in the field of *data mining* [5], which combines methods from machine learning and statistics for automatically extracting useful information from data.

Under this scenario, at first sight, it may seem that cognitive principles that inspired early machine learning research have become useless in fields such as data mining. In fact, some authors point out several problems that would arise when porting some machine learning techniques to data mining, claiming that they assume that a small, well structured, error-free dataset from which learning takes place exists. Using a database for learning may cause several problems as databases contain data generated for purposes other than learning [10]. These problems are the size of databases, both in number of instances and features, the existence of noisy and incomplete data and the need for user interaction.

Langley [12] proposes a different view pointing out that psychological modeling can still help in the design of learning systems. However, when modeling cognitive behaviors simple and noise-free datasets are needed in order to evaluate the learning processes and the results obtained. Large, high dimensional and noisy datasets like those handled in data mining would render analysis of the underlying processes and their results very complex to asses the cognitive plausibility of the methods. The goal of this paper is to show that these supposedly naive cognitive-based methods can be successfully extended to work with more complex problems, thus demonstrating that cognitive principles may be still useful in solving data mining problems.

II. Unsupervised learning and clustering

Research on inductive learning has been traditionally split into supervised and unsupervised learning. In *supervised learning*, observations are labeled by some external teacher indicating the class membership and the learning task consists in finding the characterization of each predefined class. The *unsupervised learning* task assumes that no previous information exists about the class membership of the observations, so learning systems must also find the underlying structure of the domain.

The most typical unsupervised learning task in machine learning and data mining is *clustering* [7], [11], consisting in discovering useful groups of examples and, possibly, conceptual descriptions for these groups. In psychology, the analogous task is referred to as *sorting* or *categorization*[16], where items are shown to a subject with instructions to partition them into categories.

Clustering has been extensively studied in statistics, giving raise to several well-known methods like agglomerative algorithms or optimization methods such as k-means [11]. Statistical research in clustering has often focused in data sets described by continuous features and the methods may pose some difficulties to non-experienced users, since they should be familiar with the statistical concepts involved in the clustering process to be able to tune the results. Nominal data is often found in symbolic Artificial Intelligence (AI) domains, so it is not surprising that researchers on this area developed *symbolic clustering* methods to deal with these kind of problems. Particularly, machine learning researchers have developed methods for

*The author is with the Dept. Llenguatges i Sistemes Informatics, Universitat Politecnica Catalunya, Campus Nord, Modul C6, Jordi Girona 1-3, 08034 Barcelona (Spain).E-mail: talavera@lsi.upc.es

conceptual clustering [7], [14] aiming to provide a better integration between the clustering and interpretation stages of the data analysis process.

Probably, the better known conceptual clustering system is Fisher's COBWEB[7] which in turn, constitutes an example of a design influenced by psychological concerns. COBWEB incorporates several ideas from categorization studies made by cognitive psychologists and imposes a framework based upon human learning abilities. The system addresses the clustering task as a heuristic search problem and, although it employs some statistical concepts, it is not limited by the strong assumptions required in purely statistical approaches.

III. A CASE STUDY: FEATURE SELECTION VS. SELECTIVE ATTENTION

As we discussed before, one of the problems in data mining is the size of the data sets, and particularly, the number of features. It is likely that large feature sets will contain features that are not relevant for the learning task and also will increase the complexity of learning. To solve this problem, *feature selection* methods have been developed, although mainly in the field of supervised learning.

On the other hand, in sorting experiments, psychologists have observed a tendency in humans to focus their cognitive effort on some properties when creating categories. This behavior is often referred to as *selective attention* in Cognitive Psychology and sometimes results in a very strong bias [1].

A. Selective attention: modeling human behavior

Human categorization has been always an important concern in Cognitive Psychology. Historically, one can identify two different views about how humans structure knowledge about categories. The most traditional one, is referred to as the *classical view* and assumes that categories follow strict definitions of necessary and sufficient conditions. This view imposes an all-or-none structure to categories so that all objects show the same degree of membership to these categories. However, some studies appeared to suggest that humans tend to create *familily resemblance (FR)* categories in which sorting is organized around prototypes, producing a graded membership of objects in the categories. A viewpoint accounting for these effects is the *probabilistic view*, proposing that categories are organized around probabilistic descriptions.

A number of computational models of learning are based on the FR hypothesis of category construction and have been shown to be able to reproduce several human behaviors [2], [7]. However, Ahn and Medin [1] report several experiments demonstrating that these computational models of clustering, including Fisher's COBWEB, cannot account for the behavior of people in a series of sorting experiments. These experiments explored when people were inclined to organize categories probabilistically and when they focused on individual properties. Results indicated that in some situations people were inclined to construct unidimensional (1D) categories, that is, categories described by a single dimension. None of the computational models was able to account for these results and those using probabilistic concepts tended to always form FR categories.

These results suggest that people tend to attend selectively to some of the features they observe. Computational models based upon probabilistic representations would need to incorporate some attentional mechanism in order to be able to account for this behavior.

B. Feature selection: improving data analysis

When applying machine learning technology to real world problems, accuracy is not the only important concern, but there are other important issues to consider. Among them, some of the most important are the ability to deal with a large number of features and comprehensibility. Usually, comprehensibility is measured as a function of the complexity of the results, so that both issues lead to the need of building learning systems that are able to decide which features are relevant to the learning task, that is, to perform some sort of *feature selection* [3]. Following [17] we can summarize several dimensions for evaluating the particular benefits of feature selection in clustering:

- *Performance.* The set of features used in an inductive learning task is a powerful representational bias that determines the performance of a learning system. Irrelevant features may be particularly harmful in unsupervised systems, leading the system to form wrong patterns and having an impact in prediction.
- *Efficiency in the learning task.* The number of features present in the data significantly determines the complexity clustering process, especially in hierarchical clusterings. If we apply feature selection to reduce this complexity, we should expect to obtain clusterings with at least similar performance that we would had obtained by using all the available features.
- *Efficiency in the performance task.* When using a hierarchical clustering to classify unobserved objects in order to infer unknown properties, the number of features has a strong influence in the complexity of the process in the same manner we have described above. Again, selecting an appropriate subset of features may reduce this complexity while maintaining the original performance level.
- *Comprehensibility of the results.* Reducing the number of features used in the clustering process allow to provide shorter cluster descriptions to the user. Short descriptions tend to be more readable and, hence more comprehensible.

IV. EMPIRICAL RESULTS

Although data mining and cognitive psychology may view the problem of reducing the number of terms used in the induction process under a very different light, both approaches try to essentially solve a similar question. In order to prove this point, we present two

TABLE I
THE CONTROL STRATEGY OF COBWEB.

Function Cobweb(object,root)
1) Incorporate object into the root cluster.
2) **If** root is a leaf **then**
return expanded leaf with the object.
else choose the best of the following operators:
a) Incorporate the object into the best host
b) Create a new disjunct based on the object
c) Merge the two best hosts
d) Split the best host
3) If a), c) or d) recurse on the chosen host.

TABLE II
A METHOD FOR FEATURE SELECTION BASED ON AN ORDERING SCHEME.

Let $\mathcal{A}$ be a set of features
Let τ be the feature selection threshold
Function select_features($\mathcal{A}$,τ)
compute_feature_weights($\mathcal{A}$)
$max\_w = max\{weight(A_i) \mid A_i \in \mathcal{A}\}$
return $\{A_i \mid weight(A_i) \geq max\_w \times \tau\}$

experiments using a version of COBWEB augmented with a mechanism that selects the most relevant features. The first one is aimed to fit psychological findings and the second is done from the viewpoint of data analysis. Results will show how a system that is able to model human behaviors can serve as well as a useful tool in data analysis.

A. A brief description of COBWEB

COBWEB is a hierarchical clustering system that constructs a tree from a sequence of objects. The system follows a strict *incremental* scheme, that is, it learns from each object in the sequence without reprocessing previously seen objects. An object is assumed to be a vector of nominal values V_{ij} along different features A_i. COBWEB employs *probabilistic concept* descriptions to represent the learned knowledge. In this sort of representation, in a cluster C_k, each feature value has an associated conditional probability $P(A_i = V_{ij} \mid C_k)$ reflecting the proportion of objects in C_k with the value V_{ij} along the feature A_i.

The strategy followed by COBWEB is summarized in Table I. Given an object and a current hierarchical clustering, the system categorizes the object by sorting it through the hierarchy from the root node down to the leaves. At each level, the learning algorithm evaluates the quality of the new clustering resulting from placing the object in each of the existing clusters, and the quality resulting from creating a new cluster covering the new object. In addition, the algorithm considers two more actions that can restructure the hierarchy in order to improve its quality. *Merging* attempts to combine the two sibling clusters which were identified as the two best hosts for the new object; *splitting* can replace the best host and promote its children to the next higher level. The option that yields the high quality score is selected and the procedure is recursed, considering the best host as the root in the recursive call. The recursion ends when a leaf containing only the new object is created.

In order to choose among the four available operators, COBWEB uses a cluster quality function called *category utility* defined for a partition $P = \{C_1, C_2, ..., C_n\}$ of n clusters as

$$\frac{\sum_k P(C_k) \sum_i \sum_j [P(A_i = V_{ij} \mid C_k)^2 - P(A_i = V_{ij})^2]}{n} \quad (1)$$

This function measures how much a partition P promotes inference and rewards clusters C_k that increase the predictability of feature values within C_k. By using this metric, the system should be biased towards the construction of clusters allowing accurate predictions along any unobserved features.

Several design decisions in COBWEB are influenced by results in human categorization. The use of probabilistic concepts relates to studies of *typicality effects* in categorization that could not be accounted by logical concepts. The category utility function derives from research on *basic levels*, that is, levels that humans prefer in a hierarchical classification scheme. In general, most of COBWEB assumptions are consistent with much of human learning and memory.

B. A simple attentional / feature selection method

We propose a filter method of feature selection based on an *ordering* scheme. A weight is individually computed for each feature and features are ordered according to these weights. We define a *feature selection threshold* τ in the [0,1] range such that the weight required for a feature to be selected is higher for higher τ values. Our method uses the maximum computed weight as a baseline to determine which features are selected as shown in Table II. Note that, if we assume relevances to be positive, when $\tau = 0$ there is no feature selection at all, so reducing the original algorithm to a special case of our approach.

This method can be easily incorporated into COBWEB by slightly modifying the control strategy showed in Table I. First, we need to add an additional step between steps 2 and 3 of the existing algorithm. In this step a call to the *select_features* function is performed, obtaining a subset of relevant features to be stored in the current root node. Second, at each classification step, the computation of the quality function must be modified in such a way that only the subset of relevant features stored in the current root node is used.

The weighting function we use is the one proposed by Gennari [8] in the context of his CLASSIT system, an extension of COBWEB to deal with numeric features.

Gennari refers to this measure as *salience*. He defines the relative salience of a feature as its contribution to category utility (see equation 1) in a clustering. More formally, for a given feature A_i, salience is defined as follows:

$$\frac{\sum_k P(C_k) \sum_j [P(A_i = V_{ij} \mid C_k)^2 - P(A_i = V_{ij})^2]}{n} \quad (2)$$

C. Results for psychological data

Ahn and Medin performed an experiment using four sets of exemplars varying along four dimensions. The structure of each set was different, having different degrees of between-category and within-category similarity. Each subject was asked to categorize the exemplars into two groups. The goal was to assess if subjects tended to form FR or 1D categories. They compared these results with those obtained by using different clustering approaches, including Fisher's COBWEB and found that no computational approach could fit human results.

We reproduced the experiment using the original COBWEB algorithm and the augmented version presented earlier. Some modifications were introduced, however, in order to perform a fair comparison with subjects results. First, since COBWEB has not to be told the number of clusters in the data, it can produce more than two in the top level. Subjects were explicitly asked to form only two groups, so we modified the system to force to form this number of categories at the top level. Secondly, Ahn and Medin used ordered values for some features and considered as 1D sortings those based upon only one feature or two adjacent features. As COBWEB has no knowledge about ordering and considers each value independently, we consider as 1D sortings those in which one of the categories had all the exemplars with one or two dimensions.

We tested several τ values to see which one could fit better human results. Table III shows the results for these experiments. The table includes the percentages of sorting types (family resemblance, unidimensional or other) for the four different sets. Subjects results are adapted from [1]. The final values used were 0.65 for set A, 0.75 for set B and 0.1 for sets C and D. Results show that the modified version of COBWEB is able to fit much better subject data than the plain version, perhaps with the exception of set C, in which the system tends to form more 1D sortings than subjects. An interesting result is that different τ values for different sets were selected, since it implies that different degrees of attention are needed to fit results for different category structures. Recall that higher τ values means requiring higher weights in features to not be removed. At first sight, it appears that sets with low within-category similarity such as C and D, require low τ values while sets with high within-category similarity need higher ones. We do not have an interpretation of this behavior, which remains an interesting open question.

TABLE III
RESULTS FOR PSYCHOLOGICAL DATA

Data set		FR	1D	Others
Set A	Subjects	55%	45%	0%
	Cobweb	99%	0%	0%
	Cobweb-att	49%	51%	0%
Set B	Subjects	0%	100%	0%
	Cobweb	92%	7%	1%
	Cobweb-att	1%	98%	1%
Set C	Subjects	35%	10%	55%
	Cobweb	15%	30%	55%
	Cobweb-att	18%	32%	50%
Set D	Subjects	20%	65%	15%
	Cobweb	26%	57%	17%
	Cobweb-att	22%	63%	15%

D. Results for UCI data

Using the same modified version of COBWEB than in previous subsection, we report here results adapted from [17] obtained using real world data sets from the UCI Repository. We evaluated the results under four dimensions: accuracy, efficiency in learning, efficiency in prediction and comprehensibility. In order to evaluate the efficiency gain in the learning and prediction processes, we computed the *average number of feature tests* needed to sort the instances in the training or testing set. This number is calculated by summing the total number of features involved in evaluating the category utility metric for the different clustering choices. For instance, if an observation is being clustered in a root node with three children and using a subset of n features, we need to perform $3n$ feature tests to evaluate the category utility of incorporating the observation to each of the siblings. In learning, additional feature tests are needed to evaluate creating a new cluster, merging and splitting. We think that this way of measuring efficiency give us a better empirical approximation of the complexity of the clustering process than, for instance, the average of features per node. On the other hand, we use this later measure as a measure of *comprehensibility* of the obtained clusterings, since fewer features per node indicate simpler cluster descriptions.

Table IV shows the results for the label prediction performance task. At a glance, we can observe that in all datasets accuracy can be maintained or improved while reducing the number of features per node used to an average of the 40% of the original number of features. As expected, this reduction implies an improvement in the efficiency of the system in both learning and prediction. In average, feature selection provides an efficiency improvement of about the 50% in learning and prediction.

V. GENERAL DISCUSSION

In previous sections, we have shown how an implementation of a selective mechanism that selects

TABLE IV
RESULTS FOR SEVERAL UCI DATA SETS.

Dataset	Algorithm	Accuracy	Tests learn	Tests pred	Feat./node
cleve	Cobweb	74.73 (5.05)	1619.94 (42.94)	720.40 (45.02)	13.00 (0.00)
	Cobweb-FS	76.04 (4.60)	928.64 (117.88)	363.80 (45.76)	5.22 (0.17)
crx	Cobweb	80.24 (2.89)	2226.49 (59.25)	950.22 (35.62)	15.00 (0.00)
	Cobweb-FS	80.87 (3.35)	653.60 (33.76)	211.09 (15.19)	3.91 (0.15)
horse	Cobweb	74.23 (4.60)	3108.48 (79.31)	1371.85 (126.94)	22.00 (0.00)
	Cobweb-FS	75.95 (3.85)	1548.66 (102.93)	602.92 (49.33)	8.52 (0.26)
hypo	Cobweb	97.65 (0.48)	8448.86 (248.21)	3952.57 (234.33)	25.00 (0.00)
	Cobweb-FS	97.65 (0.34)	3867.14 (229.60)	2024.41 (243.95)	18.46 (0.18)
pima	Cobweb	65.11 (2.62)	1135.25 (21.92)	470.92 (15.74)	8.00 (0.00)
	Cobweb-FS	66.06 (2.94)	571.35 (45.10)	195.13 (19.45)	3.35 (0.08)
wdbc	Cobweb	91.93 (1.55)	4287.82 (79.96)	1881.39 (70.54)	30.00 (0.00)
	Cobweb-FS	92.57 (1.20)	2249.09 (70.47)	864.46 (57.95)	11.68 (0.30)

only the most relevant feature in a clustering – categorization- - task can successfully fulfill two different goals. First, it is able to fit better psychological results in human categorization. This does not means that the method has to be considered as a computational model of human learning, rather that the results provided are consistent with human preferences or tendencies. Secondly, the same method serves the purpose of making learning and prediction more efficient and, additionally, allows more comprehensible results to be obtained. As we have mentioned, these issues are important when facing real world data analysis problems.

We think that these results suggest that the use of cognitive biases can help to find useful constraints for existing or future data mining techniques and that additional examples can be found to support this claim. One of the most representatives is the use of *incremental learning*. Incremental constraints are considered to arise in many real-world situations. Humans have the capability of modifying their conceptual schemes as new examples are observed and so learn from a stream of observations. Hence, it is not surprising that an important part of the research in unsupervised learning had focused in incremental systems. Particularly, Gennari, Langley and Fisher [9] delineate the task of *concept formation*. Concept formation systems are presented as incremental learners using a hill climbing strategy that operates under reduced search control, with low memory requirements. Another common feature of concept formation systems is their hierarchical organization of concepts, which provide a logarithmic average assimilation cost. The idea was to rapidly exploit information during learning rather than obtain an optimal result. Although the efficiency constraints present in the incremental learning approach are mainly cognitively justified, they still may be useful to deal with large amounts of data. Additionally, incremental learners provide a faster adaptation and response in the face of new data. This situation is typical in new application domains such as the design of adaptive user interfaces [13].

Probably, the most useful cognitive biases for data mining problems are going to be those related to the notion of *bounded rationality* [15], assuming that it may be desirable to reduce the assimilation cost of knowledge even if this means a decrease in quality. In fact, as opposed to most of machine learning research, data mining is not mainly concerned with accuracy, but it also stresses efficiency and comprehensibility concerns. One can view this approach as related to Simon's notion of satisfiability vs. optimality [15], claiming that heuristic AI approaches were able to give non-optimal but reasonable solutions of very complex problems that statistical optimization approaches could not provide. As discussed before, statistics and AI techniques are closer, but still it may appear somewhat surprising that some of the main concerns of an applied discipline such as data mining are reminiscent of ideas proposed thirty years ago in a context related to cognition.

Finally, it is worth stating at this point two important questions regarding the application of these ideas to real-world problems. In the first place, we have to point out that there is a strong chance that a great amount of adaptation or tuning should be necessary. For example, selective attention and feature selection face basically the same problem, but it is likely that not every attentional learning mechanism will work for every data mining problem without changes. Actually, the need of improving an existing model to solve a problem is recognized as a general requirement in the process of applying machine learning algorithms [4].

Secondly, we cannot assume that every data mining problem has a cognitive-inspired counterpart. Rather, some problems must be approached by directly using more 'pure' statistical principles, although still combined with heuristics and control strategies characteristic of AI approaches. As an example, consider the application of statistical tests in concept tree pruning strategies within the COBWEB framework [6] aimed to reduce the noise sensitivity of the system.

VI. Concluding remarks

We suggested that the role of cognitive principles in providing useful biases and/or procedures for data analysis systems may be more important that actually recognized by current research tendencies. We have discussed a particular example in unsupervised learning where the notions of selective attention in cognitive psychology and feature selection in data analysis appear to be roughly the same. Empirical evidence shows that a clustering system augmented with a mechanism that selects the most salient features can account for psychological phenomena and, at the same time, provide a significant reduction in complexity of analyzing real world data. We think that these results confirm that, as remarked by Langley [12], using knowledge of human behavior to constrain the design of learning algorithms makes good heuristic sense, not only in cognitive modeling, but in fields such as data mining as well.

We expect to see an increase in the application of ideas that can be traced back to earlier research related to cognitive concerns, basically to those based upon the *cognitive economy* principle. Of course, topics such as incremental learning are not exclusive from cognitive modeling research, but still this research will provide interesting insights and constrains to help future developments.

References

[1] W. Ahn and D. L. Medin. A two-stage model of category construction. *Cognitive Science*, (16):81–121, 1992.

[2] J. R. Anderson and M. Matessa. Explorations of an incremental, bayesian algorithm for categorization. *Machine Learning*, (9):275–308, 1992.

[3] A. L. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97:245–271, 1997.

[4] C. Brodley and Padhraic Smyth. The process of applying machine learning algorithms. In D. Aha and P. Riddle, editors, *Working Notes for Applying Machine Learning in Practice: A Workshop at the Twelfth International Machine Learning Conference*, Washington, DC, 1995.

[5] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: An overview. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 1–34. AAAI Press, Cambridge, Massachusetts, 1996.

[6] D. Fisher and P. Chan. Statistical guidance in symbolic learning. *Annals of Mathematics and Artificial Intelligence*, (2):135–148, 1990.

[7] D. H. Fisher. Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, 2:139–172, 1987.

[8] J. H. Gennari. Concept formation and attention. In *Proceedings of the Seventh Annual Conference of the Cognitive Science Society*, pages 724–728, Irvine,CA, 1991. Lawrence Erlbaum Associates.

[9] J. H. Gennari, P. Langley, and D. Fisher. Models of incremental concept formation. *Artificial Intelligence*, (40):11–61, 1989.

[10] M. Holsheimer and A. Siebes. Data mining: the search for knowledge in databases. Technical Report CS-R9406, Computer Science/Department of Algorithmics and Architecture, CWI, 1994.

[11] A. K. Jain and R. C. Dubes. *Algorithms for clustering data*. Prentice Hall, Englewood Cliffs, NJ, 1988.

[12] P. Langley. *Elements of machine learning*. Morgan Kaufmann, San Francisco, CA, 1995.

[13] P. Langley. Machine learning for adaptive user interfaces. In *Proceedings of the 21st German Annual Conference on Artificial Intelligence*, pages 53–62, 1997.

[14] R. S. Michalski and R. E. Stepp. Learning from observation: Conceptual clustering. In R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, editors, *Machine Learning: An Artificial intelligence approach*, pages 331–363. Morgan Kauffmann, Los Altos, CA, 1983.

[15] H. A. Simon. *Sciences of the artificial*. MIT Press, Cambridge,MA, 1969.

[16] E. E. Smith and D. L. Medin. *Categories and concepts*. Harvard University Press, Cambridge,MA, 1981.

[17] L. Talavera. Dynamic feature selection in incremental hierarchical clustering. In *Proceedings of the Twelfth European Conference on Machine Learning*. Springer Verlag, 2000.